Keeping Honest: Do Warnings Administered to Non-Fakers on Personality Measures Reduce Scores?

Matthew M. Vrbka

University of Minnesota

Abstract

A sample of participants (N=197), recruited from Amazon MTurk, completed an online personality assessment. Participants entered either a faking group, that promised greater rewards based on high scores, or a non-faking group, that rewarded participants the same regardless of responses. During the exam, participants were randomly assigned to a warning condition, either receiving or not receiving a warning, in order to determine the effects of warning non-fakers. Results were non-significant, but examination of interaction plots suggests non-fakers tend to lower their scores when warned.

As businesses seek to expand their prestige and productivity, more are turning to personality measures in order to hire the best possible applicants. Highhouse (2008) demonstrated personality measures, as well as other mechanical types of decision making, are better indicators of future job performance than are more traditional, clinical measures, such as unstructured interviews. However, one criticism of reliance on personality measures, as well as other types of self-report measures, is their susceptibility to manipulation and dishonest responses, or faking.

Faking can best be described as a "tendency to deliberately present oneself in a more positive manner than is accurate in order to meet the perceived demands of the testing situation" (Fan et al., 2012). It is important to note that faking is distinct from other types of response alterations, such as self-deceptive enhancement (SDE). SDE is an "unconscious response bias based on an overly positive self-image" (Fan et al., 2012). Instead, faking is a form of impression management (IM), which refers to response alterations that are intentional. Although faking remains uncommon and does not reduce the validity of personality assessments (Ones & Viswesvaran, 1998), it is nevertheless important to study, as it rears its ugly head in numerous undesirable scenarios.

By having an extremely low selection ratio, an organization may be inadvertently hiring good fakers, rather than qualified applicants, as the line between the two becomes blurry. In fact, faking is most likely to occur in more competitive organizations with lower perceived selection ratios (Robie, 2006).

These lower perceived selection ratios can be affected by a number of factors. The primary factor affecting the ratios is the number of hurdles used in the selection process, such as

multiple rounds of interviews, assessments, or screenings, and the sophistication of the predictors used within those hurdles. Companies that institute many hurdles with accurate and valid predictors would be expected to hire fewer fakers than those with few hurdles and poorly-defined predictors. Unfortunately because these methods decrease the perceived selection ratios the opposite is often true (Robie, 2006).

Another component of the perceived selection ratio is the prestige or notoriety of the company. Applicants would expect popular and successful organizations to have more rigorous methods for screening their employees, in addition to much larger applicant pools. Anecdotal information about the selection process can also affect an applicant's perceived selection ratio. For instance, if a college student has their eyes set on a prestigious, competitive graduate program, but knows none of their peers were accepted, they would perceive the school as having quite a low selection ratio.

Ultimately, an organization must set and choose a margin of error in its selection ratio with which it is comfortable. Jobs where honesty is essential, such as lawyers or police officers, may wish to set higher selection ratios to ensure an honest hire is made. Conversely, jobs that require intense skill and intelligence may not be overly concerned about a few false hires from faking, as those that fake would be expected to have a higher attrition rate. Despite this, it behooves all organizations to catch and detect faking whenever possible, in order to limit hiring deceptive applicants.

Faking can be detected via several methods of varying complexity. One of the simplest, most common, and easiest to assess is blatant extreme responding (BER). BER detects when respondents choose the most extreme of desirable options to achieve the best possible score on a measure (Levashina, 2014). In personality testing this could mean answering with 1s or 5s on

Likert scales in order to make it seem that one is very high in desirable traits, such conscientiousness, or very low in an undesirable trait, such as neuroticism.

Another type of faking detection is impression management (IM) detection. Impression management detection assesses when respondents claim to engage in desirable but very uncommon "good" behaviors, such as never telling a lie, and when respondents claim to never engage in undesirable but very common "bad" behaviors, such as stealing a pencil from work (Fan et al., 2012). This method is more difficult because it involves the analysis of the likelihood of behaviors compared to respondents' answers and requires additional measures that are unlikely to be part of a typical application process.

A final type of faking detection is the bogus statements (BS) method. In the bogus statements method assessments ask participants if they have skills, knowledge or experience that doesn't actually exist. If the participants endorse these fabricated skills, knowledge or experience they are engaging in deception (Kuncel & Borneman, 2007). What makes this method more difficult is that the bogus statements must be created by the assessors and would likely have to be tailored to specific jobs so they are perceived as being real. A bogus skill for an electrician might sound real for an electrician, but the same skill might sound like gibberish to a salesperson. This method also comes with a tradeoff, as a respondent might not have deceptive intentions when answering bogus statements. For instance, a respondent may confuse a similar-sounding bogus concept with a real concept.

Alternatively, warnings are a type of faking prevention that can be applied on a personal level. Warnings involve influencing applicants against deception and can be utilized in a variety of fashions, including detection, consequences, appeal to reason, educating, and appealing to

moral principles, to reduce the scores of fakers to a more reasonable level (Dilchert & Ones, 2012).

The benefits of warnings include their ease of implementation, low cost, and the ability to warn in real time while an applicant is taking a measurement electronically. Landers, Sackett, & Tuzinski (2011) explored this last benefit and determined that warnings administered in real time to test takers flagged as engaging in BER significantly decreased the rate of BER within the sample. Fan et al. (2012) modified this procedure by using a more rigorous flagging system and allowed flagged participants to change answers they had previously faked after the warning. They further applied this practice to both flagged and non-flagged participants, giving four groups in total: flagged with warning, flagged without warning, non-flagged with warning, and non-flagged without warning.

If a warning given to non-flagged participants does not significantly reduce their scores, compared to unwarned non-flagged participants, a warning could be given indiscriminately without any adverse effects on non-faking participants. However, Fan et al. (2012) found when non-flagged participants were warned, participants tended to overcorrect their choices, leading to somewhat lower scores than the non-flagged and unwarned group.

One possible limitation of this study is its confinement to China, which may affect the generalizability of the results. Cultural differences may account for different findings in the West. For instance, China has a collectivist culture that values and respects authority. If a test taker were to view a warning originating from an authority position, such as a researcher or hiring manager conducting the test, they might alter their responses to appease the authority figure. In the United States, an individualistic culture that values self-reliance and competition, a

non-faking test taker may disregard the warning (Triandis, Bontempo Villareal, Asai, & Lucca, 1988).

Another limitation, in both Landers et al. (2011) and Fan et al. (2012), is the reliance on BER and BS and IM scales to flag potential fakers. While both BS and IM scales offer more intense detection methods than BER, they are still susceptible to their own misinterpretations and errors. Furthermore, Kuncel, Borneman, and Kiger (2012) theorize respondents view test taking as a social interaction, and as such are unlikely to fake in the extremes in order to maintain plausibility. Similarly, some dimensions would be undesirable at an extremely high or extremely low level, for instance sociability, and as such, respondents answer somewhere in the middle when faking good on these dimensions.

To account for responding as a social interaction, and the variety of different faking methods for different dimensions, we propose a new way of "flagging" fakers: assigning them to fake implicitly. Previous lab studies have demonstrated participants are able to fake good when instructed; however, questions have been raised about the generalizability to an applicant setting (Viswesvaran & Ones, 1999). For this study, we assigned respondents to faking or non-faking groups by altering the rewards, giving individuals in the faking group an opportunity to earn a greater reward if their score is high enough. This method more closely mirrors an applicant setting, as the participants are not being explicitly instructed to fake. While it would be quite possible for fakers to exist in the non-faking group and vice versa, we would expect relatively equal overlap on both sides, in essence cancelling each other out. These groups were then assigned to be given a warning or no warning, with the warning group being allowed to change previous responses. Thus, four groups are established: fakers with warning, fakers without warning, non-fakers with warning, and non-fakers without warning.

We are especially interested in the differences between the non-faking groups, as well as testing the generalizability of Fan et al. (2012) to both collectivist and individualistic cultures. We hypothesize the non-faking group with the warning will score significantly lower than the non-faking group without the warning, supporting the generalizability of Fan et al. (2012).

Method

Participants

A total of n=200 participants was recruited for this study from Amazon Mturk. Of these 200, 45% were female. 13% were ages 18-24, 80% were 25-44, and 7% were 45-64. 6% were unemployed, and 14.5% possessed a high school diploma or GED, 25.5% had completed some college, 49% possessed a college degree, and 10% had completed graduate school. 3 of the participants either failed a trap question designed to catch speeders, or completed the survey in an unsatisfactory amount of time (under 30 seconds). As such, the amount of data-producing participants was N=197.

Materials

A modified version of the 100-item Big Five Aspect Scales (BFAS) (DeYoung, Quilty, and Peterson, 2007) was used as the personality measure. The aspect sub-dimensions of the BFAS provide more precision than is available in other standard Big Five measurement scales, with alpha coefficients ranging from .75 to .89. To tailor the scale to a workplace context, the phrase "At work, I…" was added to the start of every question. 10 questions from the 10 factors (Volatility and Withdrawal for Neuroticism, Compassion and Politeness for Agreeableness, Industriousness and Orderliness for Conscientiousness, Enthusiasm and Assertiveness for

Extraversion, and Intellect and Openness for Openness/Intellect), some of which were reverse-scored, were administered to each participant. The facet scores were then calculated for each participant and combined to form the dimension scores. A composite score for each participant was computed by averaging the dimension scores, with Neuroticism being reverse-scored. A participant scoring the absolute highest on every dimension would receive a 5 for their composite score.

Procedure

Two separate MTurk Human Intelligence Tasks (HITs) were created for this experiment. The first HIT involved a group of 100 participants completing the BFAS for a reward of $0.80. Each participant received the same reward regardless of their responses to the survey. This group was designated as the implicit non-faking group, as participants had no incentive to engage in deception.

The second HIT recruited 100 participants. Each participant completed the BFAS for a minimum of $0.80; however, participants were told that, if they scored highly enough, they could earn $1.50. Participants in this group were paid $1.50 regardless of their responses. This group was designated as the implicit faking group, as participants had an incentive to engage in deception.

The assessment was the same for both groups; however, participants in either group were randomly assigned to either a warning condition or a non-warning condition. In the non-warning condition, respondents proceeded through the exam as normal. In the warning category, regardless of their faking group affiliation, respondents were given the following warning after approximately 30 questions:

> Based on your response patterns, we have reason to believe you may be engaging in deception. Please make sure that you are genuine, credible, and true to yourself in your responses. You may go back and change any previous responses without penalty.

After being exposed to the warning, respondents in this condition were directed through the first 30 questions of the measure a second time, having the option of changing their initial responses.

All groups then completed the measure and were given a debriefing statement about the true purpose of the survey. Participants had the option of discarding their answers after reading the debriefing statement. Finally, participants were given a code to enter into the MTurk system to receive payment.

Results

Table 1 indicates the means and standard deviations of the composite score for all groups. To test our hypothesis that warned non-fakers will score lower than unwarned non-fakers, a two-way analysis of variance was conducted, as can be seen in Table 2. The analysis indicated no significant main effect for the warning, $F(1, 193)=3.46$, $p=0.26$, no significant main effect of faking or non-faking group membership, $F(1, 193)=2.07$, $p=0.39$, and no significant interaction effect, $F(1, 193)=0.05$, $p=0.81$. Table 3 contains the Type III two-way analysis of variance for the unbalanced design for the composite score. Identical tests were run on every personality

dimensions, yielding similar insignificant results. As none of the results were significant, a post-hoc test could not be practically performed.

Although the analysis indicated no significant difference between members of the warned and unwarned group and no significant difference between members of the faking and non-faking group, interaction plots were examined for composite scores and dimension scores.

Figure 1 shows the interaction plot between faking group and warning condition on composite score. Graphically, it appears there is a main effect for both faking groups and warning conditions, albeit non-significant, with unwarned fakers scoring the highest, followed, in order, by unwarned non-fakers, warned fakers, and warned non-fakers. The composite scores of unwarned non-fakers and warned fakers are similar, providing some support to the idea that warning fakers reduce their scores to a non-faking level. It is also interesting to note from the plot that the disparity between warned and unwarned non-fakers is greater than the disparity between warned and unwarned fakers.

All dimensions showed a similar interaction plot to Figure 1, with the exceptions of Extraversion (Figure 2) and Agreeableness (Figure 3). The interaction plot for Extraversion demonstrates score for unwarned non-fakers are higher than score for unwarned fakers. The interaction plot for Agreeableness indicates signs of an interaction, a main effect for warning condition, and a main effect for faking group affiliation. Here again, the disparity of non-fakers is much greater than fakers when warned.

Discussion

This experiment found no conclusive evidence that non-fakers exposed to a warning and non-fakers without a warning had different scores. However, the interaction plot for the

composite score indicated a hierarchy of sorts, similar to the results found in Fan et al. (2012), suggesting that non-fakers who receive a warning tend to lower their scores, even more than fakers who receive warnings, as demonstrated in Figure 1.

Further research is necessary to describe the true effects of warning non-fakers. This is especially important in cases of false positives, flagging a non-faker as a faker. If the composite score disparity in non-fakers truly is bigger than the disparity in fakers, a warned non-faker would not only be at a selection disadvantage by being labeled as engaging in deception, but the responder's personality score might also be drastically impacted.

Furthermore, if additional studies support these findings, a provocative question arises in the faking literature. How *meaningful* are warnings? If warnings are only applied to fakers, with a reasonably small amount of error, then the scores of fakers can successfully be brought down to that of non-fakers, and the two scores from the different groups are comparable. However, we would expect, if warning are altering the scores of fakers by bringing them closer to their true scores, warned non-fakers should be relatively unaffected. This does not appear to be the case, as the scores of warned non-fakers also decrease.

Thus, we must question what exactly warnings are doing. It is quite possible that warnings are exhibiting demand characteristics on personality measure respondents, causing all respondents to lessen the strength of their answers, simply because the test is telling them to do so. Rather than transforming the deceptive score of fakers to a value closer to their true score, it can be conjectured warnings are merely regressing scores, of fakers and non-fakers alike, to a lower value.

If this is indeed the case, faking experts should attempt to find more rigorous and precise tools in flagging potential fakers to reduce adversely impacting honest responders. Ideally, if

warnings are regressing scores, a pervasive and robust faking prevention or faking detection method would render warnings obsolete. For example, if psychometricians administered a personality measure to participants in a laboratory setting and asked participants to fake good, the researchers could find an averaged "faking best" response for every question on the measure. Those in charge of hiring decisions could then set an acceptable cut-off limit for the percentage of responses that match this ideal faking test. If the hiring manager sets the cut-off limit for this specific measure for this specific job at 90%, then those respondents that match the exact responses of the "faking best" test with 90% accuracy or more would be asked to take the test again or not continue in the hiring process. This is simply a suggestion of one of many ways the faking field can advance beyond warnings for personality measurements, and much more research is needed, on methods involving warnings and methods not involving warnings alike.

This experiment had numerous flaws and limitation, primarily the small sample size, which limited the power of our statistical analyses. Additionally, the unbalanced nature of the design, as a result of attrition and random assignment, restricted the type of analysis that could be used. In the future, a balanced research design with much larger sample sizes should be used, ideally with thousands—if not, tens of thousands—of participants.

A second limitation or source of error lies in the reward structure of the experiment. The faking group was promised $1.50 if they scores highly enough and the non-faking group was promised $0.80. The disparity between $0.80 and $1.50 might not have been high enough to prompt a participant to fake. It is also conceivable that reward differences will not cause respondents to fake, as rewards alone may not be the sole purpose a respondent has for faking. A faker in a selection context may fake in order to obtain a job and other benefits, rather than a

one-time payment. Future studies should focus on increasing this disparity, for instance using $1 and $10, to examine the legitimacy of implicit faking assignment.

A third limitation of this study, though somewhat veiled, is the medium by which it was conducted. Amazon MTurk's culture thrives off quality and accuracy as MTurk members compete for a special "Mechanical Turk Master" class of the system. To utilize these quality members for HITs, an additional fee is required. As per the Mturk website:

> "Masters are elite groups of Workers who have demonstrated accuracy on specific types of HITs on the Mechanical Turk marketplace. Workers achieve a Masters distinction by consistently completing HITs of a certain type with a high degree of accuracy across a variety of Requesters. Masters must continue to pass our statistical monitoring to remain Mechanical Turk Masters." (Amazon, 2015)

This title is determined by lottery for qualified workers and requires a suggested approval rating of 98% or greater. Thus, MTurk participants value and are rewarded on honesty. Being disapproved on even one HIT can drastically alter a participant's chances of being promoted to Master.

As a result of this culture, this study's examination of faking was not appropriate in this medium of Amazon MTurk. The researchers received several emails from participants in both the faking and non-faking groups who did not complete the test, as they found the claim that they were engaging in deception offensive. Researchers in the future are urged to be conscientious of the medium when disseminating surveys en masse. Although MTurk should likely be avoided for studies on faking, the honest responders make MTurk a tantalizing option for other experiments.

Conclusion

This study attempted to find the differences in composite BFAS scores for warned and unwarned non-fakers. While the analyses did not produce any significant results, the interaction plots of the data agreed with previous research with regards to non-fakers. Implications and possible explanations for why warnings lower scores of fakers and non-fakers were discussed. Future research should examine enhanced faking detection options and improved prevention methods.

References

Amazon.com, Inc. (2015). *Amazon Mechanical Turk Worker Web Site FAQs*. Retrieved from https://www.mturk.com/mturk/help?helpPage=worker#what_is_master_worker

Dilchert, S. Ones, D.S. (2011) Application of Preventive Strategies. In Ziegler, M., MacCann, C., & Roberts, R. (Eds.), *New perspectives on faking in personality assessment* (177-200). Oxford University Press, USA.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology*, *97*(4), 866.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333-342.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, *15*(2), 220-231.

Kuncel, N. R., Borneman, M., & Kiger, T. (2011). Innovative item response process and Bayesian faking detection methods: More questions than answers. In Ziegler, M., MacCann, C., & Roberts, R. (Eds.), *New perspectives on faking in personality assessment* (102-112). Oxford University Press, USA.

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology*, *96*(1), 202.

Levashina, J., Weekley, J. A., Roulin, N., & Hauck, E. (2014). Using Blatant Extreme Responding for Detecting Faking in High-stakes Selection: Construct validity,

relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment, 22*(4), 371-383.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human performance*, *11*(2-3), 245-269.

Robie, C., Tuzinski, K. A., & Bly, P. R. (2006). A survey of assessor beliefs and practices related to faking. *Journal of Managerial Psychology*, *21*(7), 669-681.

Triandis, H. C., Bontempo, R., Villareal, M. J., Asai, M., & Lucca, N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of personality and Social Psychology*, *54*(2), 323.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and psychological measurement*, *59*(2), 197-210.

Appendix

Table 1
Descriptive Statistics of Composite BFAS Scores for Each Group

		Composite	
Variable	N	M	SD
Main Sample	197	3.68	0.46
Faking	97	3.73	0.44
Non-Faking	100	3.34	0.47
Warning	98	3.62	0.46
No Warning	99	3.74	0.45
Faking & Warning	50	3.68	0.43
Faking & No Warning	47	3.78	0.45
Non-Faking & Warning	48	3.70	0.45
Non-Faking & No Warning	52	3.57	0.48

Note. M and *SD* represent mean and standard deviation, respectively.

Table 2
Means and standard deviations for Composite BFAS Scores in the 2x2 design

	Condition			
	No Warning		Warning	
Group	*M*	*SD*	*M*	*SD*
Faking	3.78	0.45	3.68	0.43
Non-Faking	3.70	0.45	3.57	0.48

Note. M and *SD* represent mean and standard deviation, respectively.

Table 3
Type III 2-Way ANOVA Results

Source	df	SS	F	p
Warning	1	0.27	1.30	0.26
Group	1	0.15	0.73	0.39
Warning:Group	1	0.01	0.06	0.81
Residuals	193	39.93		

Figure 1

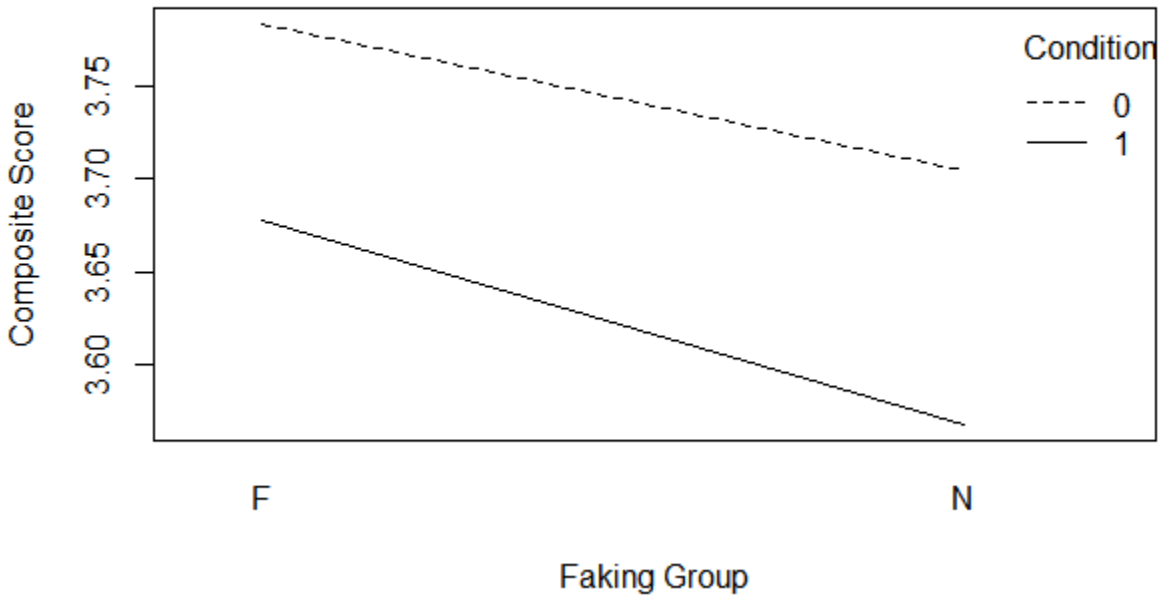

Figure 1. Interaction plot for Composite Score of Group X Condition two-way interaction.

F=fakers; N=non-fakers; 0=unwarned; 1=warned.

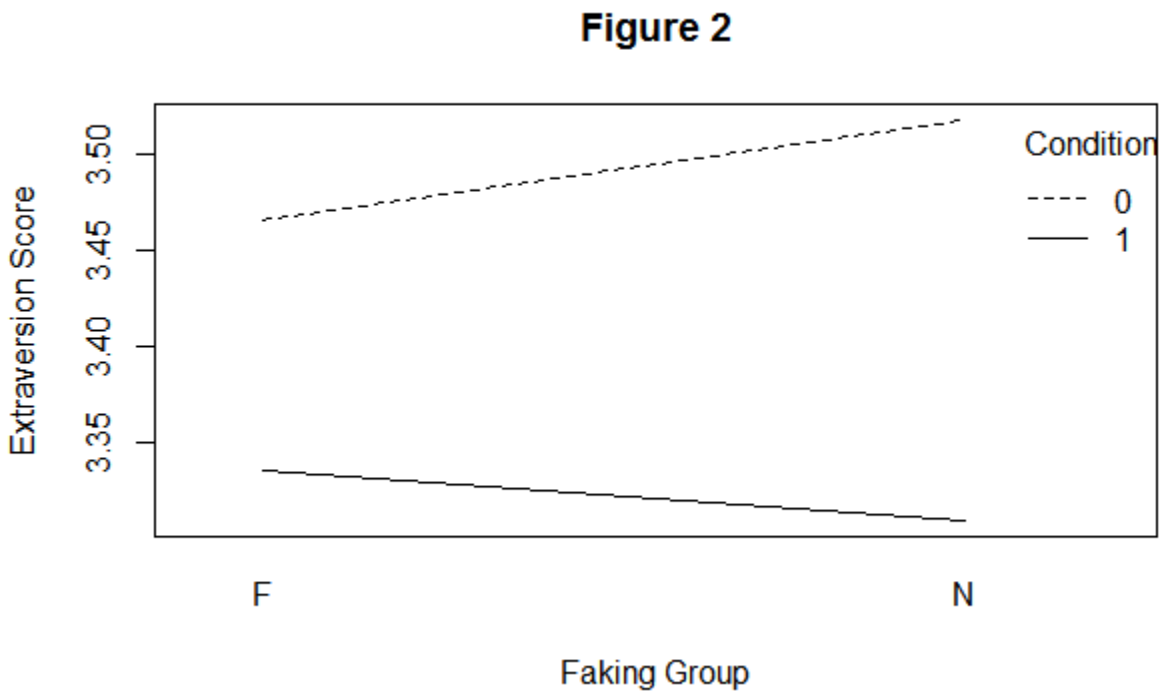

Figure 1. Interaction plot for Extraversion Score of Group X Condition two-way interaction.

F=fakers; N=non-fakers; 0=unwarned; 1=warned.

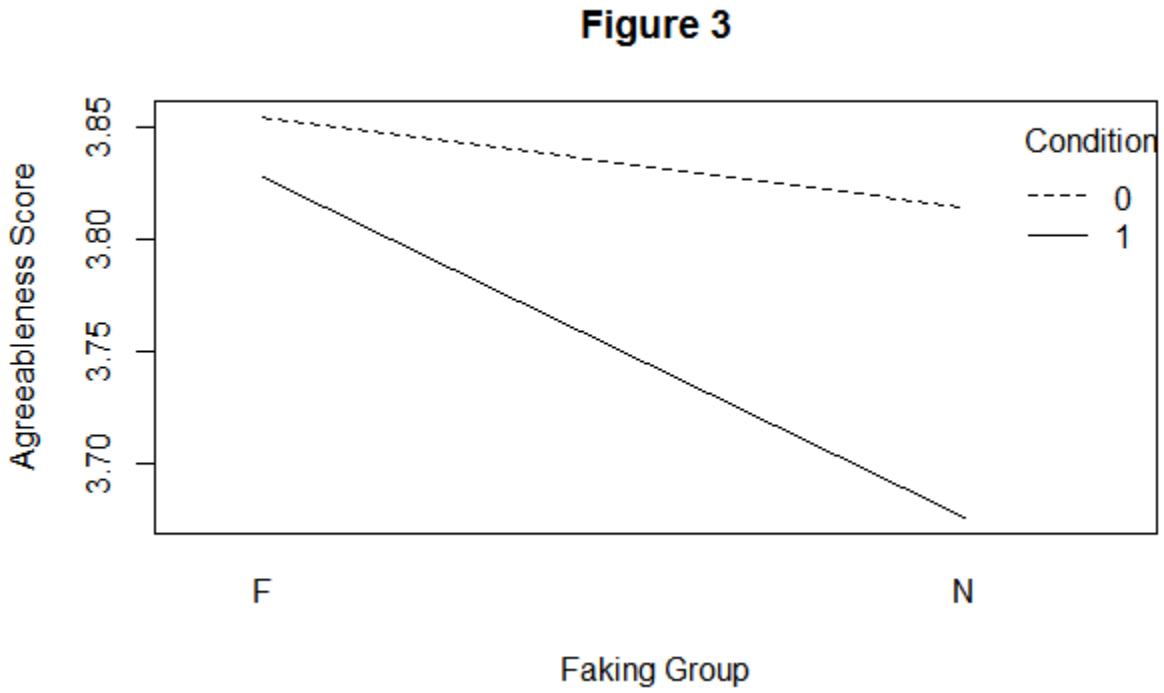

Figure 1. Interaction plot for Extraversion Score of Group X Condition two-way interaction.

F=fakers; N=non-fakers; 0=unwarned; 1=warned.